INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.862

(02/2001)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Methods for objective and subjective assessment of quality

Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs

ITU-T Recommendation P.862

(Formerly CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30 P.300
Transmission standards	Series	P.40
Objective measuring apparatus	Series	P.50 P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of quality	**Series**	**P.80** **P.800**
Audiovisual quality in multimedia services	Series	P.900

For further details, please refer to the list of ITU-T Recommendations.

ITU-T Recommendation P.862

Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs

Summary

This Recommendation describes an objective method for predicting the subjective quality of 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs. This Recommendation presents a high-level description of the method, advice on how to use it, and part of the results from a Study Group 12 benchmark carried out in the period 1999-2000. An ANSI-C reference implementation, described in Annex A, is provided in separate files and form an integral part of this Recommendation. A conformance testing procedure is also specified in Annex A to allow a user to validate that an alternative implementation of the model is correct. This ANSI-C reference implementation shall take precedence in case of conflicts between the high-level description as given in this Recommendation and the ANSI-C reference implementaion.

This Recommendation includes an electornic attachment containing an ANSI-C reference implementation of PESQ and conformance testing data.

Source

ITU-T Recommendation P.862 was prepared by ITU-T Study Group 12 (2001-2004) and approved under the WTSA Resolution 1 procedure on 23 February 2001.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

© ITU 2001

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from ITU.

CONTENTS

		Page
1	Introduction	1
2	Normative references	1
3	Abbreviations	1
4	Scope	2
5	Conventions	4
6	Overview of PESQ	4
7	Comparison between objective and subjective scores	6
7.1	Correlation coefficient	6
7.2	Residual errors	6
8	Preparation of processed speech material	7
8.1	Source material	7
8.1.1	Choice of source material	7
8.1.2	ITU-T Temporal structure and duration of source material	7
8.1.3	Filtering and level calibration	8
8.2	Addition of background noise	8
8.3	Processing through system under test	8
9	Selection of experimental parameters	8
10	Description of PESQ algorithm	9
10.1	Level and time alignment pre-processing (Figure 3)	13
10.1.1	Computation of the overall system gain	13
10.1.2	IRS filtering	13
10.1.3	Time alignment	13
10.2	Perceptual model (Figures 4a and 4b)	15
10.2.1	Precomputation of constant settings	15
10.2.2	IRS-receive filtering	15
10.2.3	Computation of the active speech time interval	15
10.2.4	Short-term Fast Fourier Transform	16
10.2.5	Calculation of the pitch power densities	16
10.2.6	Partial compensation of the original pitch power density for transfer function equalization	16
10.2.7	Partial compensation of the distorted pitch power density for time-varying gain variations between distorted and original signal	16
10.2.8	Calculation of the loudness densities	16
10.2.9	Calculation of the disturbance density	17
10.2.10	Cell-wise multiplication with an asymmetry factor	17

Page

10.2.11 Aggregation of the disturbance densities over frequency and emphasis on soft parts of the original 17
10.2.12 Zeroing of the frame disturbance for frames during which the delay decreased significantly 18
10.2.13 Realignment of bad intervals 18
10.2.14 Aggregation of the disturbance within split second intervals 18
10.2.15 Aggregation of the disturbance over the duration of the speech signal (around 10 s), including a recency factor 18
10.2.16 Computation of the PESQ score 18

Annex A – Reference implementation of PESQ and conformance testing 19

Electronic attachment: ANSI-C reference implementation of PESQ and conformance testing data.

ITU-T Recommendation P.862

Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[1]

1 Introduction

The objective method described in this Recommendation is known as "Perceptual Evaluation of Speech Quality" (PESQ). It is the result of several years of development and is applicable not only to speech codecs but also to end-to-end measurements.

Real systems may include filtering and variable delay, as well as distortions due to channel errors and low bit-rate codecs. The PSQM method as described in ITU-T P.861 (February 1998), was only recommended for use in assessing speech codecs, and was not able to take proper account of filtering, variable delay, and short localized distortions. PESQ addresses these effects with transfer function equalization, time alignment, and a new algorithm for averaging distortions over time. The validation of PESQ included a number of experiments that specifically tested its performance across combinations of factors such as filtering, variable delay, coding distortions and channel errors.

It is recommended that PESQ be used for speech quality assessment of 3.1 kHz (narrow-band) handset telephony and narrow-band speech codecs.

2 Normative references

The following ITU-T Recommendations and other reference contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

– ITU-T P.800 (1996), *Methods for subjective determination of transmission quality.*

– ITU-T P.810 (1996), *Modulated noise reference unit (MNRU).*

– ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

– ITU-T P-series Supplement 23 (1998), *ITU-T coded-speech database*.

3 Abbreviations

This Recommendation uses, the following abbreviations:

ACR Absolute Category Rating

CELP Code-Excited Linear Prediction

DMOS Degradation Mean Opinion Score

HATS Head And Torso Simulator

---

[1] This Recommendation includes an electronic attachment containing an ANSI-C reference implementation of PESQ and conformance testing data.

IRS Intermediate Reference System

LQ Listening Quality

MOS Mean Opinion Score

PCM Pulse Code Modulation

PESQ Perceptual Evaluation of Speech Quality

PSQM Perceptual Speech Quality Measure

4 Scope

Based on the benchmark results presented within Study Group 12, an overview of the test factors, coding technologies and applications to which this Recommendation applies is given in Tables 1 to 3. Table 1 presents the relationships of test factors, coding technologies and applications for which this Recommendation has been found to show acceptable accuracy. Table 2 presents a list of conditions for which the Recommendation is known to provide inaccurate predictions or is otherwise not intended to be used. Finally, Table 3 lists factors, technologies and applications for which PESQ has not currently been validated. Although correlations between objective and subjective scores in the benchmark were around 0.935 for both known and unknown data, the PESQ algorithm cannot be used to replace subjective testing.

It should also be noted that the PESQ algorithm does not provide a comprehensive evaluation of transmission quality. It only measures the effects of one-way speech distortion and noise on speech quality. The effects of loudness loss, delay, sidetone, echo, and other impairments related to two-way interaction (e.g. centre clipper) are not reflected in the PESQ scores. Therefore, it is possible to have high PESQ scores, yet poor quality of the connection overall.

Table 1/P.862 – Factors for which PESQ had demonstrated acceptable accuracy

Test factors
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates if a codec has more than one bit-rate mode
Transcodings
Environmental noise at the sending side (See Note.)
Effect of varying delay in listening only tests
Short-term time warping of audio signal
Long-term time warping of audio signal
Coding technologies
Waveform codecs, e.g. G.711; G.726; G.727
CELP and hybrid codecs ≥4 kbit/s, e.g. G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA

Applications
Codec evaluation
Codec selection
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks
NOTE – When environmental noise is present the quality can be measured by passing PESQ the clean original without noise, and the degraded signal with noise.

Table 2/P.862 – PESQ is known to provide inaccurate predictions when used in conjunction with these variables, or is otherwise not intended to be used with these variables

Test factors
Listening levels (See Note.)
Loudness loss
Effect of delay in conversational tests
Talker echo
Sidetone
Coding technologies
Replacement of continuous sections of speech making up more than 25% of active speech by silence (extreme temporal clipping)
Applications
In-service non-intrusive measurement devices
Two-way communications performance
NOTE – PESQ assumes a standard listening level of 79 dB SPL and compensates for non-optimum signal levels in the input files. The subjective effect of deviation from optimum listening level is therefore not taken into account.

Table 3/P.862 (For further study) - Factors, technologies and applications for which PESQ has not currently been validated

Test factors
Packet loss and packet loss concealment with PCM type codecs (See Note 1.)
Temporal clipping of speech (See Note 1.)
Amplitude clipping of speech (See Note 2.)
Talker dependencies
Multiple simultaneous talkers
Bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode
Network information signals as input to a codec
Artificial speech signals as input to a codec
Music as input to a codec

Table 3/P.862 (For further study) – Factors, technologies and applications for which PESQ has not currently been validated

Test factors
Listener echo
Effects/artifacts from operation of echo cancellers
Effects/artifacts from noise reduction algorithms
Coding technologies
CELP and hybrid codecs <4 kbit/s
MPEG4 HVXC
Applications
Acoustic terminal/handset testing, e.g. using HATS
NOTE 1 – PESQ appears to be more sensitive than subjects to front-end temporal clipping, especially in the case of missing words which may not be perceived by subjects. Conversely, PESQ may be less sensitive than subjects to regular, short time clipping (replacement of short sections of speech by silence). In both of these cases there may be reduced correlation between PESQ and subjective MOS. NOTE 2 – There is some evidence to suggest that PESQ is able to account for amplitude clipping, but only four conditions are known to have been included (in two 50-condition experiments) in the validation database described in clause 7.

5 Conventions

Subjective evaluation of telephone networks and speech codecs may be conducted using listening-only or conversational methods of subjective testing. For practical reasons, listening-only tests are the only feasible method of subjective testing during the development of speech codecs, when a real-time implementation of the codec is not available. This Recommendation discusses an objective measurement technique for estimating subjective quality obtained in listening-only tests, using listening equipment conforming to the IRS or modified IRS receive characteristics.

Most information on the performance of PESQ is from ACR listening quality (LQ) subjective experiments. This Recommendation should therefore be considered to relate primarily to the ACR LQ opinion scale.

6 Overview of PESQ

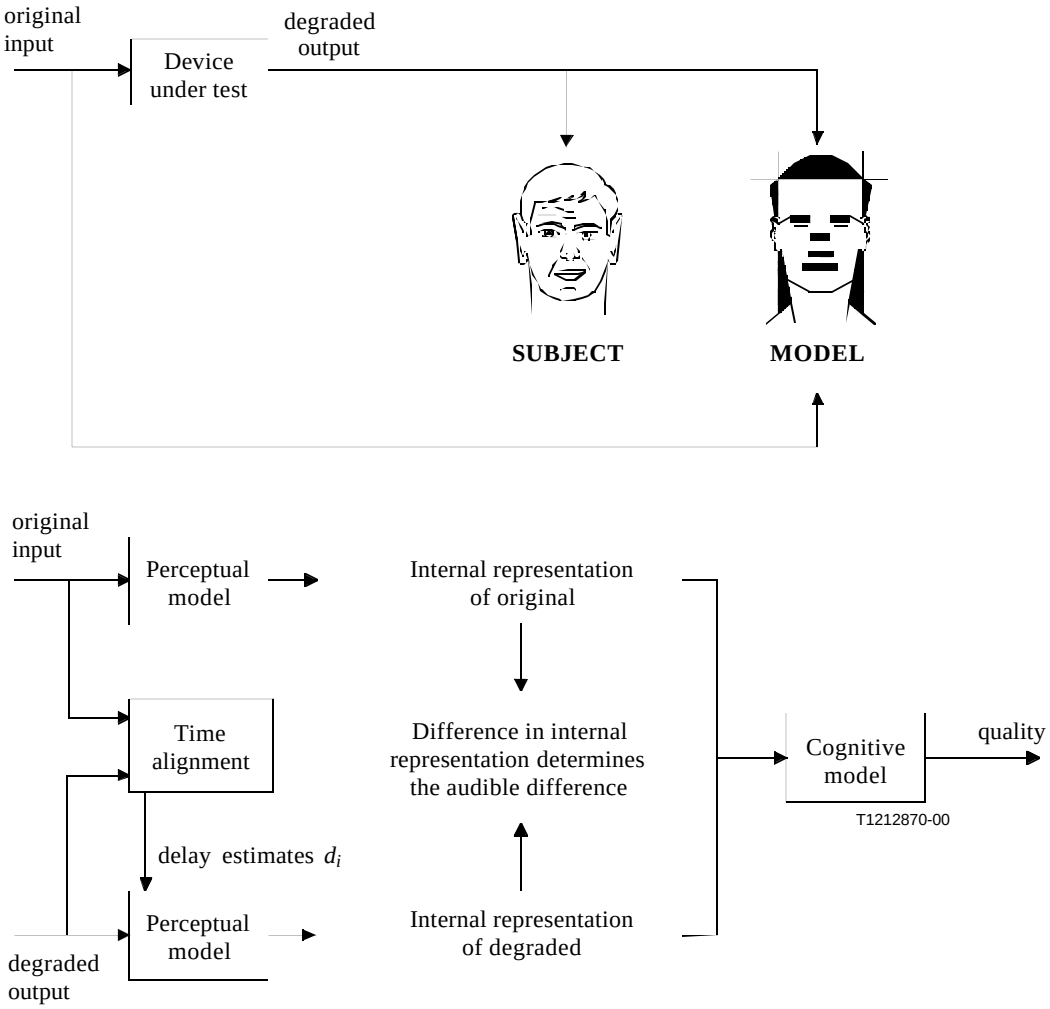
PESQ compares an original signal X(t) with a degraded signal Y(t) that is the result of passing X(t) through a communications system. The output of PESQ is a prediction of the perceived quality that would be given to Y(t) by subjects in a subjective listening test.

In the first step of PESQ a series of delays between original input and degraded output are computed, one for each time interval for which the delay is significantly different from the previous time interval. For each of these intervals a corresponding start and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. The algorithm can handle delay changes both during silences and during active speech parts.

Based on the set of delays that are found PESQ compares the original (input) signal with the aligned degraded output of the device under test using a perceptual model, as illustrated in Figure 1. The key

to this process is transformation of both the original and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). This is achieved in several stages: time alignment, level alignment to a calibrated listening level, time-frequency mapping, frequency warping, and compressive loudness scaling.

The internal representation is processed to take account of effects such as local gain variations and linear filtering that may – if they are not too severe – have little perceptual significance. This is achieved by limiting the amount of compensation and making the compensation lag behind the effect. Thus minor, steady-state differences between original and degraded are compensated. More severe effects, or rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. This allows a small number of quality indicators to be used to model all subjective effects. In PESQ, two error parameters are computed in the cognitive model; these are combined to give an objective listening quality MOS. The basic ideas which are used in PESQ are described in the bibliography references [1] … [5].



NOTE – A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the device under test with the imput, using alignment information as derived from the time signals in the time alignment module.

Figure 1/P.862 – Overview of the basic philosophy used in PESQ

7 Comparison between objective and subjective scores

Subjective votes are influenced by many factors such as the preferences of individual subjects and the context (the other conditions) of the experiment. Thus, a regression process is necessary before a direct comparison can be made. The regression must be monotonic so that information is preserved, and it is normally used to map the objective PESQ score onto the subjective score. A good objective quality measure should have a high correlation with many different subjective experiments if this regression is performed separately for each one, and in practice, with PESQ, the regression mapping is often almost linear, using a MOS-like scale.

A preferred regression method for calculating the correlation between the PESQ score and subjective MOS, which was used in the validation of PESQ, uses a 3rd-order polynomial constrained to be monotonic. This calculation is performed on a per study basis. In most cases, condition MOS is the chosen performance metric, so the regression should be performed between condition MOS and condition-averaged PESQ scores. A condition should at least use four different speech samples. The result of the regression is an objective MOS score in that test. In order to be able to compare objective and subjective scores the subjective MOS scores should be derived from a listening test that is carried out according to ITU-T P.830.

7.1 Correlation coefficient

The closeness of the fit between PESQ and the subjective scores may be measured by calculating the correlation coefficient. Normally this is performed on condition averaged scores, after mapping the objective to the subjective scores. The correlation coefficient is calculated with Pearson's formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

In this formula, x_i is the condition MOS for condition i, and $\bar{x}$ is the average over the condition MOS values, x_i, y_i is the mapped condition-averaged PESQ score for condition i, and $\bar{y}$ is the average over the predicted condition MOS values y_i.

For 22 known ITU benchmark experiments, the average correlation was 0.935. For an agreed set of eight experiments used in the final validation – experiments that were unknown during the development of PESQ – the average correlation was also 0.935.

7.2 Residual errors

The regression mapping removes any systematic offset between the objective scores and the subjective MOS, minimizing the mean square of the residual errors:

$$e_i = x_i - y_i$$

Various measures may be applied to the residual errors to give an alternative view of the closeness of objective scores to subjective MOS. For example, the histogram of the absolute residual errors $|e_i|$ provides a quick view of how frequently errors of different magnitudes occur.

For 22 known ITU benchmark experiments, the average residual error distribution showed that the absolute error was less than 0.25 MOS (±0.25 on a 5-point scale) for 69.2% of the conditions and less than 0.5 MOS. For an agreed set of 7 experiments used in the final validation, experiments that were unknown during the development of PESQ, the absolute error was less than 0.25 MOS (±0.25 on a 5-point scale) for 72.3% of the conditions and less than 0.5 MOS (±0.5 on a 5-point scale) for 91.1% of the conditions.

8 Preparation of processed speech material

It is important that test signals for use with PESQ are representative of the real signals carried by communications networks. Networks may treat speech and silence differently and coding algorithms are often highly optimized for speech – and so may give meaningless results if they are tested with signals that do not contain the key temporal and spectral properties of speech. Further pre-processing is often necessary to take account of filtering in the send path of a handset, and to ensure that power levels are set to an appropriate range.

8.1 Source material

8.1.1 Choice of source material

At present, all official performance results for PESQ relate to experiments conducted using the same natural speech recordings in both the subjective and objective tests. The use of artificial speech signals and concatenated real speech test signals is recommended only if they represent the temporal structure (including silent intervals) and phonetic structure of real speech signals.

Artificial speech test signals can be prepared in several ways. A concatenated real speech test signal may be constructed by concatenating short fragments (e.g. one second) of real speech while retaining a representative structure of speech and silence. Alternatively, a phonetic approach may be used to produce a minimally redundant artificial speech signal which is representative of both the temporal and phonetic structure of a large corpus of natural speech [6]. Test signals should be representative of both male and female talkers. In preliminary tests, high quality artificial speech and concatenated real speech both showed good results with PESQ. In these tests the objective scores for the test signals in each condition served as a prediction for the subjective condition MOS values. This approach makes it possible to determine the quality of the system under test with the least possible effort. The subject is for further study.

If natural speech recordings are used, the guidelines given in clause 7/P.830 should be followed, and it is recommended that a minimum of two male talkers and two female talkers be used for each testing condition. If talker dependency is to be tested as a factor in its own right, it is recommended that more talkers be used: eight male, eight female and eight children. Currently, ITU-T P.862 is not validated for talker dependency.

8.1.2 ITU-T Temporal structure and duration of source material

Test signals should include speech bursts separated by silent periods, to be representative of natural pauses in speech. As a guide, 1 to 3 s is a typical duration of a speech burst, although this does vary considerably between languages. Certain types of voice activity detectors are sensitive only to silent periods that are longer than 200 ms. Speech should be active for between 40 per cent and 80 per cent of the time, though again this is somewhat language dependent.

Most of the experiments used in calibrating and validating PESQ contained pairs of sentences separated by silence, totalling 8 s in duration; in some cases three or four sentences were used, with slightly longer recordings (up to 12 s). Recordings made for use with PESQ should be of similar length and structure.

Thus, if a condition is to be tested over a long period, it is most appropriate to make a number of separate recordings of around 8 to 20 s of speech and process each file separately with PESQ. This has additional benefits: if the same original recording is used in every case, time variations in the quality of the condition will be very apparent; alternatively, several different talkers and/or source recordings can be used, allowing more accurate measurement of talker or material dependence in the condition.

Note that the non-linear averaging process in PESQ means that the average score over a set of files will not usually equal the score of a single concatenated version of the same set of files.

8.1.3 Filtering and level calibration

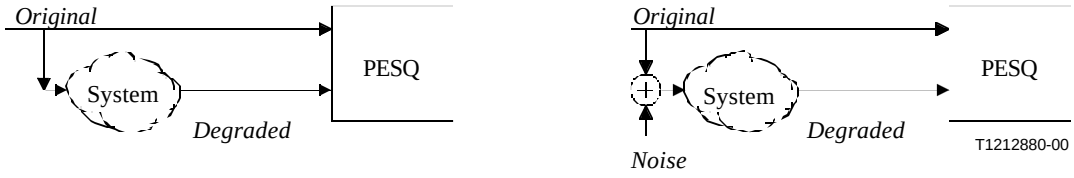
Signals should be passed through a filter with appropriate frequency characteristics to simulate sending frequency characteristics of a telephone handset, and level-equalized in the same manner as real voices. ITU-T recommends the use of the modified Intermediate Reference System (IRS) sending frequency characteristic as defined in Annex D/P.830. Level alignment to an amplitude that is representative of real traffic should be performed in accordance with 7.2.2/P.830.

In some cases the measurement system used (for example, a 2-wire analogue interface) may introduce significant level changes. These should be taken into account to ensure that the signal passed into the network is at a representative level.

The prepared source material after handset (send) filtering and level alignment is normally used as the original signal for PESQ.

8.2 Addition of background noise

It is possible to use PESQ to assess the quality of systems carrying speech in the presence of background or environmental noise (e.g. car, street, etc). Noise recordings should be passed through an appropriate filter similar to the modified IRS sending characteristic – this is especially important for low-frequency signals such as car noise which are heavily attenuated by the handset filter – and then level aligned to the desired level for the test. For PESQ to take account of the subjective disturbance in an ACR context, due to the noise as well as any coding distortions, the original signal used with PESQ should be clean, but the noise should be added before the signals are passed to the system under test. PESQ is validated for noisy input signals. The process of noise addition is shown in Figure 2b).



a) Testing with clean speech

b) Testing with noisy speech

Figure 2/P.862 – Methods for testing quality with and without environmental noise

8.3 Processing through system under test

The source signal should be processed as appropriate for the system under test. It is desirable to avoid any further distortion by unnecessary quantization, amplitude clipping, or resampling. The preferred format for storing original and degraded signals is 8 kHz sample rate, 16-bit linear PCM. PESQ was validated on both 8 and 16 kHz sampling rate.

9 Selection of experimental parameters

The effects of various quality factors on the performance of the codec or system can be examined in subjective and/or objective assessment. ITU-T P.830 provides guidance on subjectively assessing the following quality factors:

1) speech input levels to a codec;

2) listening levels in subjective experiments;

3) talkers (including multiple simultaneous talkers);

4) errors in the transmission channel between an encoder and a decoder;

5) bit rates if a codec has more than one bit-rate mode;

6) transcodings;

7) bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode;

8) environmental noise in the sending side;

9) network information signals as input to a codec;

10) music as input to a codec.

PESQ allows assessment of many of these quality factors (1, 4, 5, 6 and 8).

NOTE 1 – Objective measurement for quality factors other than those specifically noted as applicable in this Recommendation is still under study. Therefore, these factors should be measured only after the accuracy of an objective measure is verified in conjunction with subjective tests conforming to ITU-T P.830.

In addition to the codec conditions, ITU-T P.830 recommends the use of reference conditions in subjective tests. These conditions are necessary to facilitate the comparison of subjective test results from different laboratories or from the same laboratory at different times. Also, when expressing the objective test results in terms of equivalent-Q values, reference conditions using the narrow-band Modulated Noise Reference Unit (MNRU) as specified in ITU-T P.810 should be tested.

NOTE 2 – Including other standard codecs such as G.711 64-kbit/s PCM, G.726 32-kbit/s ADPCM, G.728 16-kbit/s LD-CELP, and G.729 8-kbit/s CS-ACELP as well as MNRU in objective quality measurement may assist in comparing the performance of the system under test with standardized codecs.

Detailed explanations of these experimental parameters are found in ITU-T P.830.

10 Description of PESQ algorithm

Because many of the steps in PESQ are quite algorithmically complex, a description is not easily expressed in mathematical formulae. This description is textual in nature and the reader is referred to the C source code for a detailed description. Figures 3, 4a and 4b give an overview of the algorithm in the form of a block diagram: Figure 3 for the alignment, 4a for the core of the perceptual model and 4b for the final determination of the PESQ score. For each of the blocks a high-level description is given.

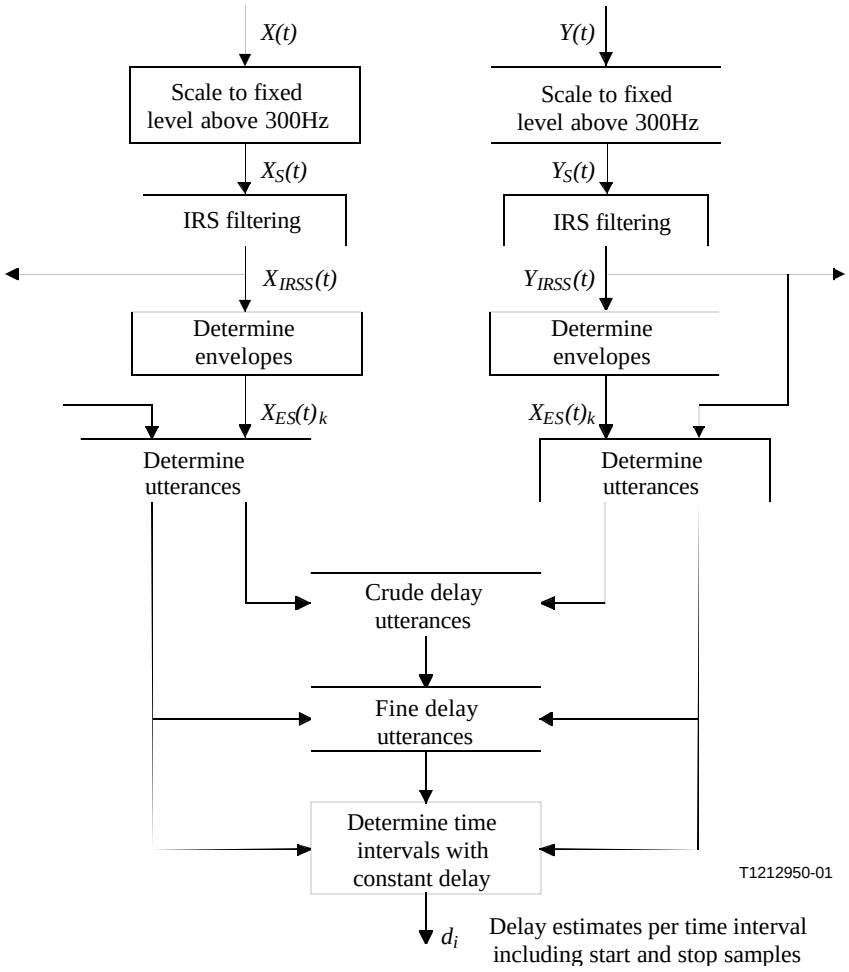
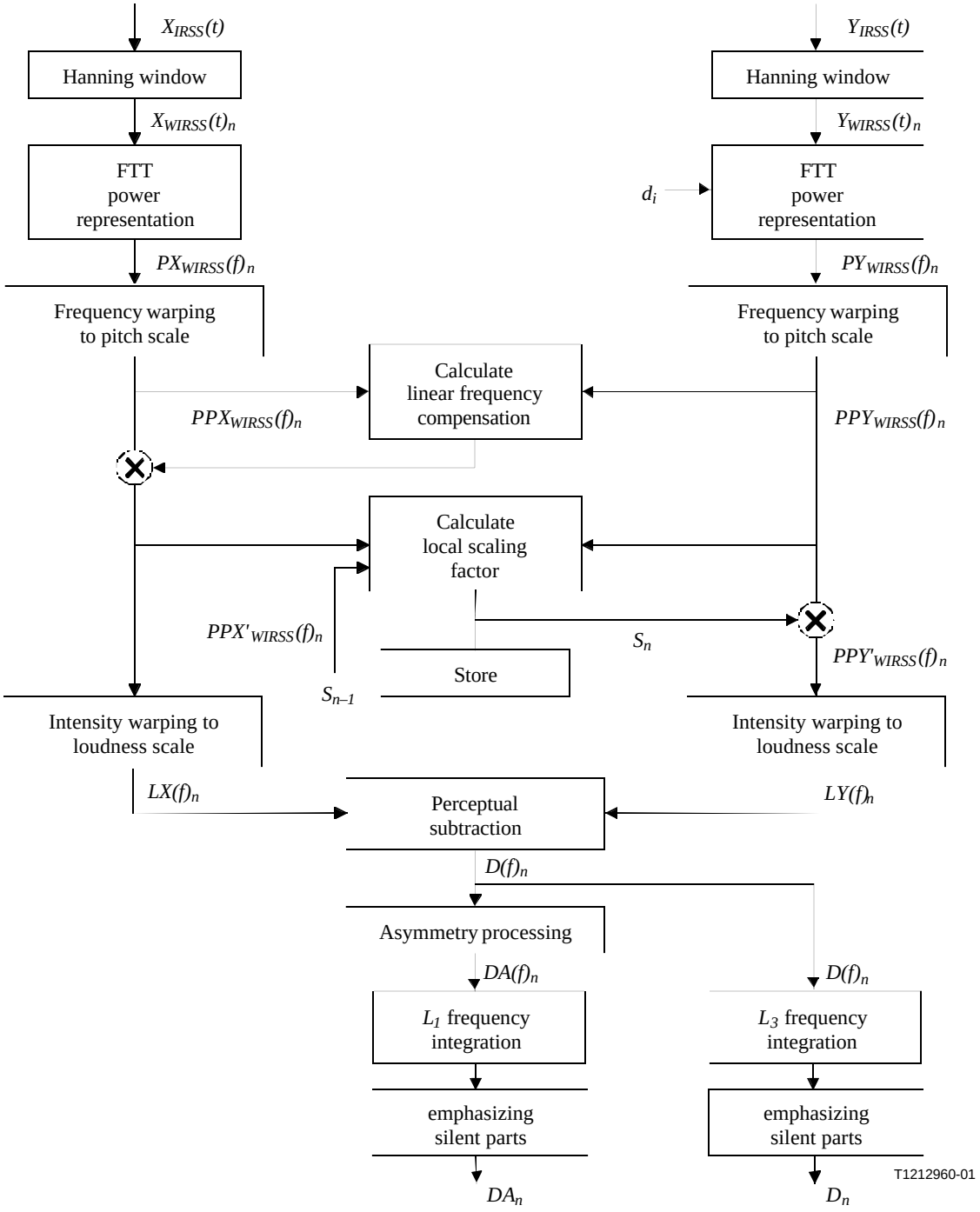
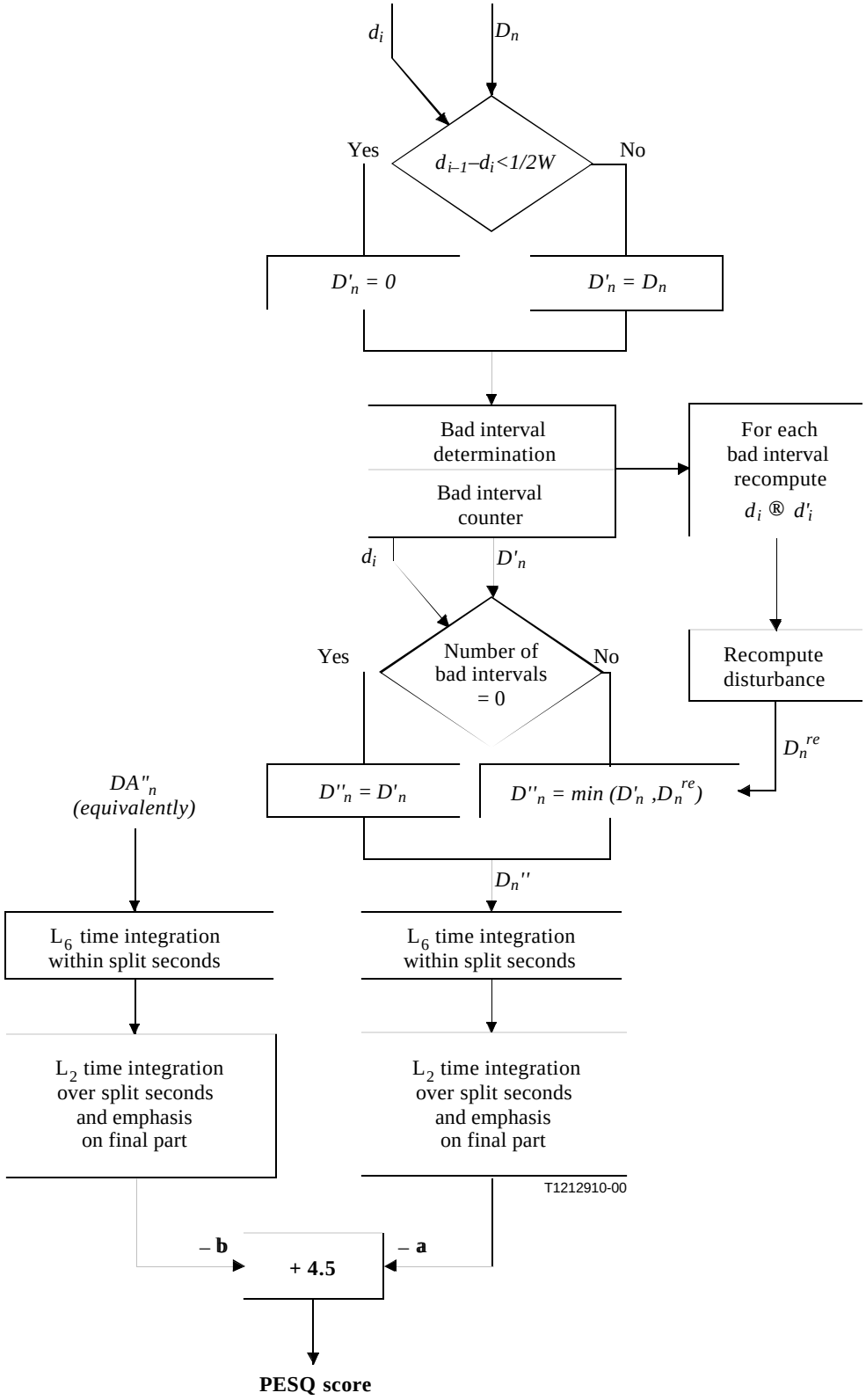


Figure 3/P.862 – Overview of the alignment routine used in PESQ to determine the delay per time interval d_i



NOTE – The distortions per frame D_n and DA_n have to be aggregated over time (index$_n$) to obtain the final disturbances (see Figure 4b where also the realignment of the degraded signal is given).

Figure 4a/P.862 – Overview of the perceptual model



NOTE – After realignment of the bad intervals, the distortions per frame D''_n and DA''_n are integrated over time and mapped to the PESQ score. W is the FTT window length in samples.

Figure 4b/P.862 – Overview of the perceptual model

10.1 Level and time alignment pre-processing (Figure 3)

10.1.1 Computation of the overall system gain

The gain of the system under test is not known *a priori* and may vary considerably, for example depending on whether an ISDN connection or an analogue 2-wire interface was used for measurement. Furthermore, there is no single calibrated level that the original signal will be stored at. Thus it is necessary to level align both the original $X(t)$ and degraded signal $Y(t)$ to the same, constant power level. PESQ assumes that the subjective listening level is a constant 79 dB SPL at the ear reference point (see 8.1.2/P.830). The level alignment algorithm in PESQ proceeds as follows:

- Filtered versions of the original and degraded signal are computed. The filter blocks all components under 250 Hz, is flat until 2000 Hz and then falls off with a piecewise linear response through the following points: {2000 Hz, 0 dB}, {2500 Hz, –5 dB }, {3000 Hz,–10 dB}, {3150 Hz, –20 dB}, {3500 Hz, –50 dB}, {4000 Hz and above, –500 dB}. These filtered versions of the signals are only used in this computation of the overall system gain.
- The average value of the squared filtered original speech samples and filtered degraded speech samples are computed.
- Different gains are calculated and applied to align both the original $X(t)$ and degraded speech signal $Y(t)$ to a constant target level resulting in the scaled versions $X_S(t)$ and $Y_S(t)$ of these signals.

10.1.2 IRS filtering

It is assumed that the listening tests were carried out using an IRS receive or a modified IRS receive characteristic in the handset. A perceptual model of the human evaluation of speech quality must take account of this to model the signals that the subjects actually heard. Therefore IRS-like receive filtered versions of the original speech signal and degraded speech signal are computed.

In PESQ this is implemented by a FFT over the length of the file, filtering in the frequency domain with a piecewise linear response similar to the (unmodified) IRS receive characteristic (ITU-T P.830), followed by an inverse FFT over the length of the speech file. This results in the filtered versions $X_{IRSS}(t)$ and $Y_{IRSS}(t)$ of the scaled input and output signals $X_S(t)$ and $Y_S(t)$. A single IRS-like receive filter is used within PESQ irrespective of whether the real subjective experiment used IRS or modified IRS filtering. The reason for this approach was that in most cases the exact filtering is unknown, and that even when it is known the coupling of the handset to the ear is not known. It is therefore a requirement that the objective method be relatively insensitive to the filtering of the handset.

The IRS filtered signals are used both in the time alignment procedure and the perceptual model.

10.1.3 Time alignment

The time alignment routine provides time delay values to the perceptual model to allow corresponding signal parts of the original and degraded files to be compared. This alignment process takes a number of stages:

- envelope-based delay estimation using entire original and degraded signals;
- division of original signal into a number of subsections known as utterances;
- envelope-based delay estimation on utterances;
- fine correlation/histogram-based identification of delay to nearest sample on utterances;
- splitting utterances and realigning the time intervals to search for delay changes during speech;

- after the perceptual model, identifying and realigning long sections of large errors to search for alignment errors.

10.1.3.1 Envelope-based alignment

The envelopes $X_{ES}(t)_k$ and $Y_{ES}(t)_k$ are calculated from the scaled original and degraded signals $X_S(t)$ and $Y_S(t)$. The envelope is defined as LOG (MAX($E(k)$/*Ethresh, 1*)), where $E(k)$ is the energy in 4 ms frame k and *Ethresh* is the threshold of speech determined by a voice activity detector. Cross-correlation of the envelopes for the original and degraded signals is used to estimate the crude delay between them, with an approximate resolution of 4 ms.

10.1.3.2 Fine time alignment

Because perceptual models are sensitive to time offsets, it is necessary to calculate a sample-accurate delay value. This is computed as follows:

- 64 ms frames (75 per cent overlapping) are Hann windowed and cross-correlated between original and degraded signals, after the envelope-based alignment is performed.
- The maximum of the correlation, to the power 0.125, is used as a measure of the confidence of the alignment in each frame. The index of the maximum gives the delay estimate for each frame.
- A histogram of these delay estimates, weighted by the confidence measure, is calculated. The histogram is then smoothed by convolution with a symmetric triangular kernel of width 1 ms.
- The index of the maximum in the histogram, combined with the previous delay estimate, gives the final delay estimate.
- The maximum of the histogram, divided by the sum of the histogram before convolution with the kernel, gives a confidence measure between 0 (no confidence) and 1 (full confidence).

The result after fine time alignment has been performed is a delay value and a delay confidence for each utterance, taking account of delay changes during silent periods. Along with the known start and end points of each utterance this allows the delay of each frame to be identified in the perceptual model.

10.1.3.3 Utterance splitting

Delay changes during speech are tested by splitting and realigning time intervals in each utterance. Envelope-based alignment is performed to compute an estimate of the delay of each part, then fine time alignment is performed to identify the delay and confidence of each part. The splitting process is repeated at several different points within each utterance and the split that produces greatest confidence is identified. If this gives greater confidence than the alignment without a split, and the two parts have significantly different delay, the utterance is divided accordingly. The test is applied recursively to each part after a split has taken place to test for further delay changes.

In this way, delay changes both during speech and during silence are accounted for and the delay per time interval (d_i) together with the matched start and stop samples are calculated. The number of time intervals is determined by the number of delay changes.

10.1.3.4 Perceptual realignment

After the perceptual model has been applied, sections that have very large disturbance (greater than a threshold value) are identified and realigned by cross-correlation. This step improves the model's accuracy with a small number of hard-to-align files where delay changes are not correctly identified by the previous time alignment process. The way this implemented is given in 10.2.13.

10.2 Perceptual model (Figures 4a and 4b)

The perceptual model of PESQ is used to calculate a distance between the original and degraded speech signal (PESQ score). As discussed in clause 7, this may be passed through a monotonic function to obtain a prediction of subjective MOS for a given subjective test. The PESQ score is mapped to a MOS-like scale, a single number in the range of –0.5 to 4.5, although for most cases the output range will be between 1.0 and 4.5, the normal range of MOS values found in an ACR listening quality experiment.

10.2.1 Precomputation of constant settings

Certain constants values and functions are pre-computed. For those that depend on the sample frequency, versions for both 8 and 16 kHz sample frequency are stored in the program.

10.2.1.1 FFT window size depending on the sample frequency (8 or 16 kHz)

In PESQ the time signals are mapped to the time-frequency domain using a short-term FFT with a Hann window of size 32 ms. For 8 kHz this amounts to 256 samples per window and for 16 kHz the window counts 512 samples while adjacent frames are overlapped by 50%.

10.2.1.2 Absolute hearing threshold

The absolute hearing threshold $P_0(f)$ is interpolated to get the values at the center of the Bark bands that are used. These values are stored in an array and are used in Zwicker's loudness formula.

10.2.1.3 The power scaling factor

There is an arbitrary gain constant following the FFT for time-frequency analysis. This constant is computed from a sine wave of a frequency of 1000 Hz with an amplitude at 29.54 (40 dB SPL) transformed to the frequency domain using the windowed FFT over 32 ms. The (discrete) frequency axis is then converted to a modified Bark scale by binning of FFT bands. The peak amplitude of the spectrum binned to the Bark frequency scale (called the "pitch power density") must then be 10 000 (40 dB SPL). The latter is enforced by a postmultiplication with a constant, the power scaling factor S_p.

10.2.1.4 The loudness scaling factor

The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After binning to the modified Bark scale, the intensity axis is warped to a loudness scale using Zwicker's law, based on the absolute hearing threshold. The integral of the loudness density over the *Bark* frequency scale, using a calibration tone at 1000 Hz and 40 dB SPL, must then yield a *value of* 1 Sone. The latter is enforced by a postmultiplication with a constant, the loudness scaling factor S_l.

10.2.2 IRS-receive filtering

As stated in 10.1.2, it is assumed that the listening tests were carried out using an IRS receive or a modified IRS receive characteristic in the handset. The necessary filtering to the speech signals is already applied in the pre-processing.

10.2.3 Computation of the active speech time interval

If the original and degraded speech file start or end with large silent intervals, this could influence the computation of certain average distortion values over the files. Therefore, an estimate is made of the silent parts at the beginning and end of these files. The sum of five successive absolute sample values must exceed 500 from the beginning and end of the original speech file in order for that position to be considered as the start or end of the active interval. The interval between this start and end is defined as the active speech time interval. In order to save computation cycles and/or storage size, some computations can be restricted to the active interval.

10.2.4 Short-term Fast Fourier Transform

The human ear performs a time-frequency transformation. In PESQ this is implemented by a short-term FFT with a window size of 32 ms. The overlap between successive time windows (frames) is 50 per cent. The power spectra – the sum of the squared real and squared imaginary parts of the complex FFT components – are stored in separate real valued arrays for the original and degraded signals. Phase information within a single Hann window is discarded in PESQ and all calculations are based on only the power representations $PX_{WIRSS}(f)_n$ and $PY_{WIRSS}(f)_n$.

The start points of the windows in the degraded signal are shifted over the delay. The time axis of the original speech signal is left as is. If the delay increases, parts of the degraded signal are omitted from the processing, while for decreases in the delay parts are repeated.

10.2.5 Calculation of the pitch power densities

The Bark scale reflects that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark does not exactly follow the values given in the literature. The resulting signals are known as the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$.

10.2.6 Partial compensation of the original pitch power density for transfer function equalization

To deal with filtering in the system under test, the power spectrum of the original and degraded pitch power densities are averaged over time. This average is calculated over speech active frames only using time-frequency cells whose power is more than 1000 times the absolute hearing threshold. Per modified Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum. The maximum compensation is never more than 20 dB. The original pitch power density $PPX_{WIRSS}(f)_n$ of each frame n is then multiplied with this partial compensation factor to equalize the original to the degraded signal. This results in an inversely filtered original pitch power density $PPX'_{WIRSS}(f)_n$.

This partial compensation is used because severe filtering can be disturbing to the listener. The compensation is carried out on the original signal because the degraded signal is the one that is judged by the subjects in an ACR experiment.

10.2.7 Partial compensation of the distorted pitch power density for time-varying gain variations between distorted and original signal

Short-term gain variations are partially compensated by processing the pitch power densities frame by frame. For the original and the degraded pitch power densities, the sum in each frame n of all values that exceed the absolute hearing threshold is computed. The ratio of the power in the original and the degraded files is calculated and bounded to the range $[3 \cdot 10^{-4}, 5]$. A first-order low-pass filter (along the time axis) is applied to this ratio. The distorted pitch power density in each frame, n, is then multiplied by this ratio, resulting in the partially gain compensated distorted pitch power density $PPY'_{WIRSS}(f)_n$.

10.2.8 Calculation of the loudness densities

After partial compensation for filtering and short-term gain variations, the original and degraded pitch power densities are transformed to a Sone loudness scale using Zwicker's law [7].

$$LX(f)_n = S_l \cdot \left(\frac{P_0(f)}{0.5} \right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_0(f)} \right)^{\gamma} - 1 \right]$$

with $P_0(f)$ the absolute threshold and S_l the loudness scaling factor from 10.2.1.4.

Above 4 Bark, the Zwicker power, γ, is 0.23, the value given in the literature. Below 4 Bark, the Zwicker power is increased slightly to account for the so-called recruitment effect. The resulting two-dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called loudness densities.

10.2.9 Calculation of the disturbance density

The signed difference between the distorted and original loudness density is computed. When this difference is positive, components such as noise have been added. When this difference is negative, components have been omitted from the original signal. This difference array is called the raw disturbance density.

The minimum of the original and degraded loudness density is computed for each time-frequency cell. These minima are multiplied by 0.25. The corresponding two-dimensional array is called the mask array. The following rules are applied in each time-frequency cell:

- If the raw disturbance density is positive and larger than the mask value, the mask value is subtracted from the raw disturbance.
- If the raw disturbance density lies in between plus and minus the magnitude of the mask value, the disturbance density is set to zero.
- If the raw disturbance density is more negative than minus the mask value, the mask value is added to the raw disturbance density.

The net effect is that the raw disturbance densities are pulled towards zero. This represents a dead zone before an actual time frequency cell is perceived as distorted. This models the process of small differences being inaudible in the presence of loud signals (masking) in each time-frequency cell. The result is a disturbance density as a function of time (window number n) and frequency, $D(f)_n$.

10.2.10 Cell-wise multiplication with an asymmetry factor

The asymmetry effect is caused by the fact that when a codec distorts the input signal it will in general be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different percepts, the input signal and the distortion, leading to clearly audible distortion [2]. When the codec leaves out a time-frequency component, the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable. This effect is modelled by calculating an asymmetrical disturbance density $DA(f)_n$ per frame by multiplication of the disturbance density $D(f)_n$ with an asymmetry factor. This asymmetry factor equals the ratio of the distorted and original pitch power densities raised to the power of 1.2. If the asymmetry factor is less than 3, it is set to zero. If it exceeds 12, it is clipped at that value. Thus, only those time-frequency cells remain, as non-zero values, for which the degraded pitch power density exceeded the original pitch power density.

10.2.11 Aggregation of the disturbance densities over frequency and emphasis on soft parts of the original

The disturbance density $D(f)_n$ and asymmetrical disturbance density $DA(f)_n$ are integrated (summed) along the frequency axis using two different Lp norms and a weighting on soft frames (having low loudness):

$$D_n = M_n \sqrt[3]{\sum_{f=1,\ldots Number\ of\ Bark\ bands} \left(|D(f)_n| W_f \right)^3}$$

$$DA_n = M_n \sum_{f=1,\ldots \text{Number of Bark bands}} \left(\left| DA(f)_n \right| W_f \right)$$

with M_n a multiplication factor, *1/(power of original frame plus a constant)*$^{0.04}$, resulting in an emphasis of the disturbances that occur during silences in the original speech fragment, and W_f a series of constants proportional to the width of the modified Bark bins. After this multiplication the frame disturbance values are limited to a maximum of 45. These aggregated values, D_n and DA_n, are called frame disturbances.

10.2.12 Zeroing of the frame disturbance for frames during which the delay decreased significantly

If the distorted signal contains a decrease in the delay larger than 16 ms (half a window), the repeat strategy as mentioned in 10.2.4 is modified. It was found to be better to ignore the frame disturbances during such events in the computation of the objective speech quality. As a consequence frame disturbances are zeroed when this occurs. The resulting frame disturbances are called D'_n and DA'_n.

10.2.13 Realignment of bad intervals

Consecutive frames with a frame disturbance above a threshold are called bad intervals. In a minority of cases the objective measure predicts large distortions over a minimum number of bad frames due to incorrect time delays observed by the pre-processing. For those so-called bad intervals, a new delay value is estimated by maximizing the cross-correlation between the absolute original signal and absolute degraded signal adjusted according to the delays observed by the pre-processing. When the maximal cross-correlation is below a threshold, it is concluded that the interval is matching noise against noise and the interval is no longer called bad, and the processing for that interval is halted. Otherwise, the frame disturbance for the frames during the bad intervals is recomputed and, if it is smaller replaces the original frame disturbance. The result is the final frame disturbances D''_n and DA''_n that are used to calculate the perceived quality.

10.2.14 Aggregation of the disturbance within split second intervals

Next, the frame disturbance values and the asymmetrical frame disturbance values are aggregated over split second intervals of 20 frames (accounting for the overlap of frames: approx. 320 ms) using L_6 norms, a higher p value as in the aggregation over the speech file length. These intervals also overlap 50 per cent and no window function is used.

10.2.15 Aggregation of the disturbance over the duration of the speech signal (around 10 s), including a recency factor

The split second disturbance values and the asymmetrical split second disturbance values are aggregated over the active interval of the speech files (the corresponding frames) now using L_2 norms. The higher value of p for the aggregation within split-second intervals as compared to the lower p value of the aggregation over the speech file is due to the fact that when parts of the split seconds are distorted, that split second loses meaning, whereas if a first sentence in a speech file is distorted, the quality of other sentences remains intact.

10.2.16 Computation of the PESQ score

The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value. The range of the PESQ score is –0.5 to 4.5, although for most cases the output range will be a listening quality MOS-like score between 1.0 and 4.5, the normal range of MOS values found in an ACR experiment.

Bibliography

[1] BEERENDS (J.G.), STEMERDINK (J.A.): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation, *J. Audio Eng. Soc.*, Vol. 42, No. 3, pp. 115-123, March 1994.

[2] BEERENDS (J.G.): Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, *Speech Quality Assessment*, Workshop papers, Bochum, pp. 1-9, November 1994.

[3] BEERENDS (J.G.): Measuring the quality of speech and music codecs, an integrated psychoacoustic approach, *98th AES Convention*, pre-print No. 3945, 1995.

[4] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.): Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain, *IEE Proceedings – Vision, Image and Signal Processing*, 141 (3), 203-208, June 1994.

[5] RIX (A.W.), REYNOLDS (R.), HOLLIER (M.P.): Perceptual measurement of end-to-end speech quality over audio and packet-based networks, *106th AES Convention*, pre-print No. 4873, May 1999.

[6] HOLLIER (M.P.), HAWKSFORD (M.O.), GUARD (D.R.): Characterisation of communications systems using a speech-like test stimulus, *Journal of the AES*, 41 (12), 1008-1021, December 1993.

[7] ZWICKER (E), FELDTKELLER (R.): Das Ohr als Nachrichtenempfänger, *S. Hirzel Verlag*, Stuttgart 1967.

ANNEX A

Reference implementation of PESQ and conformance testing

List of files provided for the ANSI-C reference implementation

The ANSI-C reference implementation of PESQ is contained in the following text files which are provided in the `source` sub-directory of the CD-ROM distribution:

- dsp.c *Basic DSP routines*
- dsp.h *Header file for dsp.c*
- pesq.h *General header file*
- pesqdsp.c *PESQ DSP routines*
- pesqio.c *File input/output*
- pesqmain.c *Main program*
- pesqmod.c *PESQ high-level model*
- pesqpar.h *PESQ perceptual model definitions*

The ANSI-C reference implementation is provided in separate files and forms an integral part of this Recommendation. The ANSI-C reference implementation shall take precedence in case of conflicts between the high-level description as given in this Recommendation and the ANSI-C reference implementation.

List of files provided for conformance validation

The conformance validation process described below makes reference to the following files, which are provided in the `conform` sub-directory of the CD-ROM distribution:

- supp23.txt *File pairs and PESQ scores for conformance validation to Supplement 23*
- supp23.bat *Batch script to assist with validation to Supplement 23*
- voipref.txt *File pairs and PESQ scores for conformance validation with variable delay*
- voipref.bat *Batch script to assist with validation with variable delay*
- or105.wav or109.wav or114.wav or129.wav or134.wav or137.wav
  or145.wav or149.wav or152.wav or154.wav or155.wav or161.wav
  or164.wav or166.wav or170.wav or179.wav or221.wav or229.wav
  or246.wav or272.wav dg105.wav dg109.wav dg114.wav dg129.wav
  dg134.wav dg137.wav dg145.wav dg149.wav dg152.wav dg154.wav
  dg155.wav dg161.wav dg164.wav dg166.wav dg170.wav dg179.wav
  dg221.wav dg229.wav dg246.wav dg272.wav
  u_am1s01.wav u_am1s02.wav u_am1s03.wav
  u_am1s01b1c1.wav u_am1s01b1c7.wav u_am1s02b1c9.wav
  u_am1s01b1c15.wav u_am1s03b1c16.wav u_am1s03b1c18.wav
  u_am1s01b2c1.wav u_am1s02b2c4.wav u_am1s02b2c5.wav
  u_am1s03b2c5.wav u_am1s03b2c6.wav u_am1s03b2c7.wav
  u_am1s01b2c8.wav u_am1s03b2c11.wav u_am1s02b2c14.wav
  u_af1s01b2c16.wav u_af1s03b2c16.wav u_af1s02b2c17.wav
  u_af1s03b2c17.wav u_am1s03b2c18.wav u_af1s01.wav
  u_af1s02.wav u_af1s03.wav

  Speech files provided for validation with variable delay.

These speech files are in Wave format (16-bit linear PCM, Intel byte ordering, 44-byte header), at 8 kHz sample rate.

Conformance validation

The conformance of an implementation of PESQ to this Recommendation must be validated as follows: The validation is based on a comparison between the reference implementation (the ANSI-C code described in this annex) and the user's implementation of PESQ. This allows the user to conduct a detailed comparison of individual variables by examining differences on any level required.

For an implementation to be deemed to conform to this Recommendation, it must pass two mandatory tests comparing the PESQ score with that given by the reference implementation on the following databases:

- ITU-T P-series Supplement 23 (1736 file pairs);
- variable delay conditions (40 file pairs).

Comparison with ITU-T P-series Supplement 23

In this comparison, all ten experiments as released with ITU-T P-series Supplement 23 are used on a file-by-file basis. The original and degraded file names, and the PESQ score given by the reference implementation, are provided in the file supp23.txt that is supplied with the PESQITU ANSI-C code.

An implementation passes this test when the absolute difference in PESQ score compared to the reference implementation is less than 0.05 in **all** files.

This conformance test is mandatory for all implementations.

Supplement 23 to the P-series Recommendations can be obtained separately from the ITU.

Comparison with variable delay files

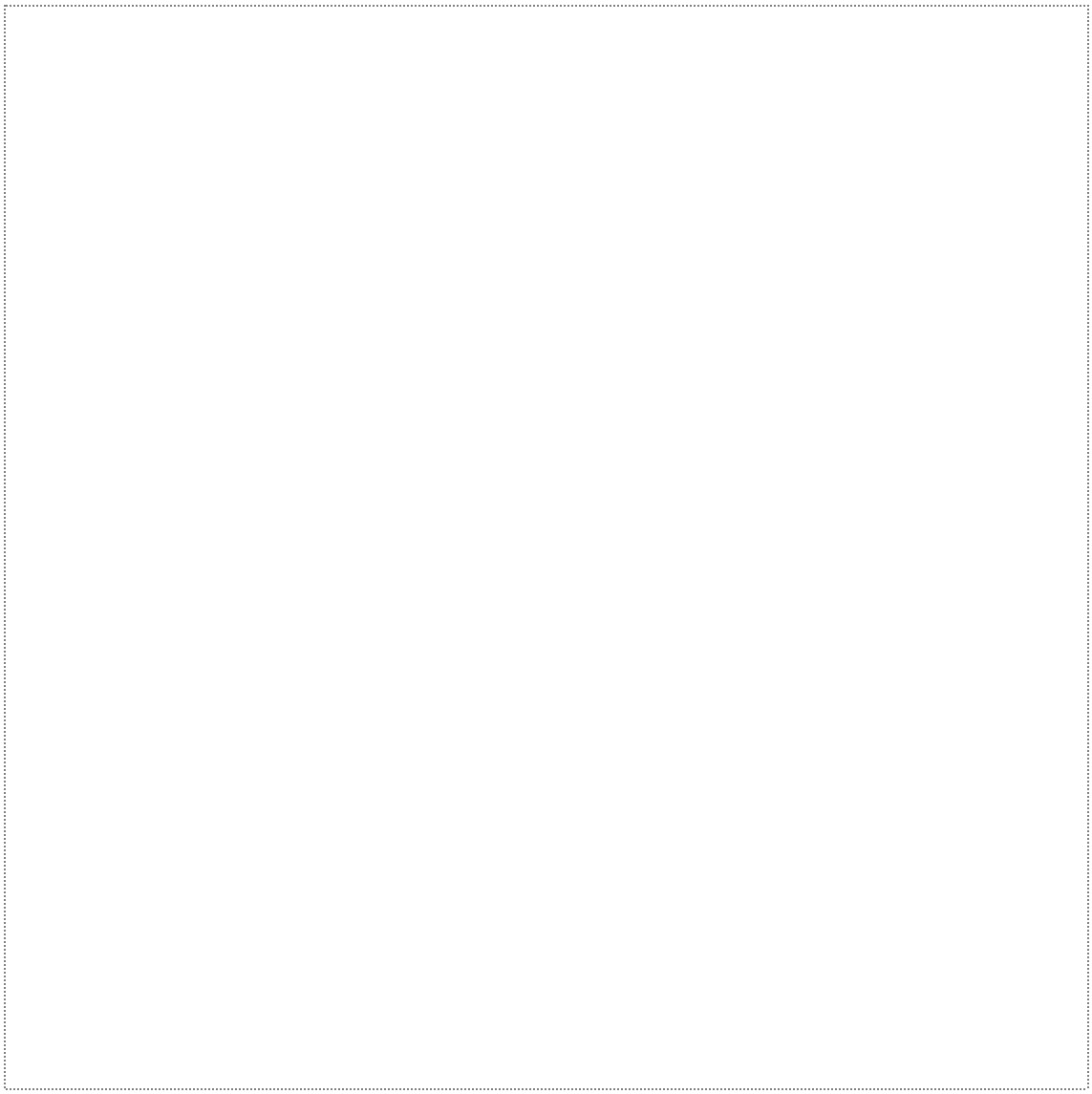
A composite database has been constructed from 40 conditions (file pairs) from two subjective tests covering real and simulated VoIP connections that exhibit time-varying delay. Twenty-three of these file pairs also trigger the bad interval realignment process. The original and degraded file names for each file pair, and the PESQ score given by the reference implementation, are provided in the file voipref.txt that is supplied with the ANSI-C code.

An implementation passes this test when the absolute difference in PESQ score compared to the reference implementation is less than 0.05 in 39 of the 40 file pairs. A single file pair is allowed to have an absolute difference in PESQ score of less than 0.5. This may be any one of the 40 file pairs.

This conformance test is mandatory for all implementations.

Additional comparisons

To prevent implementors from specifically tailoring an algorithm to conform to requirements for the files described above, a further test is available. An implementation of PESQ that conforms to this Recommendation must in at least 95% of cases give an output score that is within 0.05 of the PESQ score given by the ANSI-C reference implementation.



SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communications
Series Y	Global information infrastructure and Internet protocol aspects
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland